



07 - 03 - 00

A

EXPRESS MAIL LABEL NO. EL563154951US

DOCKET NO. ARC9-2000-0046-US1

Assistant Commissioner for Patents  
Washington, D. C. 20231

Sir:

Transmitted herewith for filing is the patent application of:

INVENTORS: Reiner KRAFT and Jussi P. MYLLYMAKI

TITLE: *SYSTEM AND METHOD FOR ENHANCED BROWSER-BASED WEB  
CRAWLING*



In connection with this application, the following are enclosed:

- 23 Pages of Specification, Claims and Abstract
- 20 Claims
- 9 Sheets of Drawings (FIGS. 1-8)
- XX Declaration, Power of Attorney
- XX Assignment to: International Business Machines Corporation

The fee has been calculated as shown below. (Small entity fees indicated in parentheses.)

For	Number Filed		Number Extra	Rate Large (Small)	Basic Fee \$690 (\$345)
Total Claims	20	20	0	\$18 (\$9)	0
Independent Claims	3	3	0	\$78 (\$39)	0
Multiple Dependent Claims				\$270 (\$135)	0
Assignment Recording Fee				\$40	40
TOTAL FEE:					\$ 730

XX The Commissioner is hereby authorized to charge Deposit Account No. 09-0441 in the amount of \$730. A duplicate copy of this sheet is enclosed.


XX The Commissioner is hereby authorized to charge payments of (1) any additional filing fees required under 37 CFR 1.16, and/or (2) any patent application processing fees under 37 CFR 1.17 associated with this application or credit any overpayment to Deposit Account No. 09-0441.

SEND CORRESPONDENCE TO:

Respectfully submitted,

FLEIT, KAIN, GIBBONS, GUTMAN  
& BONGINI, P.L.  
4400 N. Federal Highway, Suite 32  
Boca Raton, FL 33431  
(561)417-9477  
(561)417-3844 Fax

BY:

  
Jon A. Gibbons  
Reg. No. 37,333

PATENT

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of: : Atty Docket: ARC9-2000-0046-US1  
Reiner KRAFT *et al.* : APPLICATIONS BRANCH  
Serial No. (not yet assigned) :  
Filed: HEREWITH :  
FOR: SYSTEM AND METHOD FOR ENHANCED BROWSER-BASED WEB CRAWLING

CERTIFICATE OF EXPRESS MAIL MAILING

"Express Mail" Mailing Label No. **EL563154951US**  
Date of Deposit: **June 30, 2000**

Box Patent Application  
Assistant Commissioner for Patents  
Washington, D.C. 20231

SIR:

I hereby certify that

<u>X</u>	Application Transmittal
<u>X</u>	Specification, Claims, Abstract
<u>X</u>	1 set of 9 sheets of drawings
<u>X</u>	Declaration and Power of Attorney
<u>X</u>	Assignment
<u>X</u>	Return postcard

are being deposited with the United States Postal Service "Express Mail Post Office to Addressee" service under 37 CFR 1.10 on the date indicated above and are addressed to:

Box Patent Application  
Assistant Commissioner for Patents  
Washington, D.C. 20231

6/30/00  
Date of Deposit

Kathleen Smith  
Name of person mailing papers  
Kathleen Smith  
Signature

# SYSTEM AND METHOD FOR ENHANCED BROWSER-BASED WEB CRAWLING

(Provided for Examination Reference Purposes)

PARTIAL WAIVER OF COPYRIGHT .....	1
CROSS-REFERENCE TO RELATED APPLICATIONS .....	1
FIELD OF THE INVENTION .....	1
BACKGROUND OF THE INVENTION .....	1
SUMMARY OF THE INVENTION .....	4
BRIEF DESCRIPTION OF THE FIGURES .....	6
DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS .....	7
Glossary of Terms Used in this Disclosure .....	7
Exemplary Embodiment-System Architecture for Enhanced Browser Based Crawler .....	10
URL POOL 302a .....	11
Visited Pool 304a .....	11
Pool Manager 306a .....	11
Page Gatherer 308a .....	12
Page Renderer 310a .....	12
Resource Cache 312a .....	13
Page Extractor 314a .....	13
Page Summarizer 316a .....	14
Discussion of Hardware and Software Implementation Options .....	14
CLAIMS .....	16
ABSTRACT .....	23

**EXPRESS MAIL NO.: EL563154951US**

**DATE MAILED:** June 30, 2000

**PATENT**

**INVENTORS:** Reiner KRAFT  
Jussi P. MYLLYMAKI

## **SYSTEM AND METHOD FOR ENHANCED BROWSER-BASED WEB CRAWLING**

### **PARTIAL WAIVER OF COPYRIGHT**

5  
10  
All of the material in this patent application is subject to copyright protection under the copyright laws of the United States and of other countries. As of the first effective filing date of the present application, this material is protected as unpublished material. However, permission to copy this material is hereby granted to the extent that the copyright owner has no objection to the facsimile reproduction by anyone of the patent documentation or patent disclosure, as it appears in the United States Patent and Trademark Office patent file or records, but otherwise reserves all copyright rights whatsoever.

### **CROSS-REFERENCE TO RELATED APPLICATIONS**

Not Applicable

### **15 FIELD OF THE INVENTION**

This invention relates to Internet Search Technology and E-Commerce and more particularly to an improved method and apparatus for web crawling.

### **BACKGROUND OF THE INVENTION**

20 Database systems store enormous amounts of information that can be accessed by users for identification and retrieval of valuable documents that contain data, text, audio and video information. A typical example of a database system 100 is shown in FIG. 1. Information processing units 101a to 101n can be any of the following: personal computers

**EXPRESS MAIL NO.: EL563154951US**

(DOS, WINDOWS or Macintosh, Linux machines), workstations, a client, a dumb terminal or equivalent. Hub processing units 102a to 102y can be any of the following: a server, a master, a database controller or equivalent. Network (100) can be any of the following: a token ring network, a star network, a telecommunication switching network, a local area network (LAN), a wide area network (WAN), a corporate intranet, the Internet or equivalent. Information processing units 101a to 101n are in communication with hub processing units 102a to 102y via network 100. The sharing of data across network 100 is accomplished by computer search programs 103a to 103x operating in conjunction with the hub processing units 102a to 102y. The search programs can be located on the hub processing units themselves or on another processing units that are not shown. In addition, a user employs a graphical user interface (GUI) 104a to 104n that permits him or her to submit search queries across network 100 to the hub processing units.

Upon reception of the search query, the hub processing units forward the request to the search programs 103a to 103x for completion of the transaction. As is well known, search programs provide Boolean Operators (AND, OR NOT) to help build more sophisticated queries in order to narrow down the search result set. These Boolean Operators are used to provide the various permutations to the search programs 103a to 103x which uses these to locate pertinent documents. Once in possession of the search query, the search programs compare the requested search parameters against documents stored in databases 105a to 105z. Finding words or phrases that compare favorably with the search query, the search programs return a list of relevant documents to the information processing units 101a to 101n as well as library information such as type of document, location and highlighted words or phrases indicating the flags that caused the search program to retrieve the particular document. Finally, the search results are loaded into the graphical user interface GUI 104a to 104n for the user's review.

The search programs 103a to 103x used to return the search results in FIG.1 are commonly referred to as "web crawlers". Today's web crawling methodologies are already able to retrieve heterogeneous, static content from the World Wide Web (WWW). However, as more and more designers use dynamically generated content in their web-

**EXPRESS MAIL NO.: EL563154951US**

based documents, existing crawling techniques are not always capable of retrieving the data correctly. Known Enhanced Crawling architectures are able to simulate user interaction; thus, these enable automatic crawling of web sites that dynamically generate their data and associate data with session information.

5 Referring now to the flow diagram 200 of FIG.2, a typical web crawler performs two main operations in order to execute the crawling process; namely, the access - retrieval of a document (202) and then the analysis phase of the document, also called the summarization process (204). Whereas today's web crawler might be able to access a dynamically generated document correctly (e.g., a document generated through Active  
10 Server Pages, Perl Script or an equivalent), the summarization process will fail or produce flawed results if the document itself contains executable client side software code. The reason for this is that the client side software code (e.g., JavaScript, VBScript, or equivalent) is targeted to be executed and interpreted within a web browser's scripting engine. Eventually the code will be replaced with content, or the code produces content. Web designers often make use of this feature to dynamically create content on the client  
15 side; examples of this can include computation results which are originated based on some user input, or specific text based on the client's web browser version used or some other such equivalent. More generally, dynamic documents rely on a web browser's capabilities to:

- 20
- a) retrieve additional documents (206) as needed (frames, in-line images, audio, video, applets, or equivalents) or required;
  - b) execute client side script (208) and code (JavaScript or equivalents);
  - c) furnish a fault tolerant HTML filter to recognize various HTML standards and  
25 interpret HTML markup errors; unscramble content that a web designer has purposefully scrambled in order to thwart crawling and other programmatic analysis methods; thereby produce a final HTML markup (210); and
  - e) integrate all these previously obtained results to render (212) the document for presentation to a user (214).

**EXPRESS MAIL NO.: EL563154951US**

As one can see the information unit 101a to 101n side web browsing 104a to 104n process can become very complicated and convoluted. That's the reason why it's not a trivial task to implement a decent web browser. Further, there are additional problems involved in this implementation. As an example, a web browser has to achieve fault tolerance in regard to the underlying HTML used to create a document. First, there are many different HTML versions and standards currently available. Second, human error is introduced into the document when individuals do not correctly compose HTML markup. They forget brackets, use the wrong syntax, arguments, parameters or other such errors that necessitate a fault-tolerant browser. Therefore, there is a need for a fault-tolerant web crawler that does not fail when summarizing dynamic documents.

Further, today's web designers often make intensive use of images and image maps to represent text data in documents. Some of these documents consist only of images and the images themselves contain all the textual data and other information in the document. However, standard web crawlers will not be able to summarize such a document. Therefore, there is a need for a web crawler that can interpret and summarize textual and other information contained within the body of a web-based image document.

In summary, a web browser has to execute a complicated algorithmic process in order to eliminate the problems previously described; this complex algorithmic process enables the browser to present and render a document in the manner that the web document composer intended it to be displayed. A web browser's functionality is similar to that of a multi-tasking management component which has to coordinate several tasks to yield an effective end product. The web browser must coalesce information from a variety of sources to produce the final HTML code which will be rendered and displayed.

**SUMMARY OF THE INVENTION**

This invention pioneers an enhanced crawling mechanism and technique called "Enhanced Browser Based Web Crawling". It permits the fault-tolerant gathering of dynamic data documents on the World Wide Web (WWW). The Enhanced Browser Based Web Crawler technology of this invention is implemented by incorporating the

**EXPRESS MAIL NO.: EL563154951US**

intricate functionality of a web browser into the crawler engine so that documents are properly analyzed. Essentially, the Enhanced Browser Based Crawler acts similarly to a web browser after retrieving the initially requested document. It then loads additional or included documents as needed or required (e.g., inline-frames, frames, images, applets, audio, video, or equivalents). The Crawler then executes client side script or code and produces the final HTML markup. This final HTML markup is ordinarily used for the rendering for user presentation process. However, unlike a web browser this invention does not render the composed document for viewing purposes. Rather it analyzes or summarizes it, thereby extracting valuable metadata and other important information contained within the document.

In another embodiment, this invention introduces the integration of Optical Character Recognition (OCR) techniques into the crawler architecture. The reason for this is to enable the web crawler summarization process to properly summarize image content (e.g., GIF, JPEG or an equivalent) without errors. Since today's web designers often make intensive use of images and image maps to represent textual and other data in documents, it is imperative that a web crawler be capable of retrieving and summarizing images and image maps that contain textual or other data types in a fault-free manner.

Using the Enhanced Browser Based Crawler of this invention to enhance existing document gathering and analysis introduces significant advantages over the prior art. First, the quality of the extracted metadata is dramatically improved. This is due to the fact that the summarization of a document is based on the whole and complete document as it was designed by the document's author; the static heterogeneous data as well as the problematic dynamic data is processed fault-free and integrated into the metadata. A standard web crawler is not able to compose this type of highly dynamic and distributed document that includes dynamic information such as client side script, applets, or their equivalents. Secondly, the integration of optical character recognition (OCR) techniques into the document analysis and summarization process enables the retrieval of textual data from images or image maps. This text can be analyzed and added to the document summary.



**EXPRESS MAIL NO.: EL563154951US**

Overall the enhanced browser crawling technique described in the invention produces a higher quality of metadata, because it can integrate and analyze information which cannot be obtained from standard crawling techniques. As a result, a search engine provider utilizing this invention, is able to provide a virtually fault-free search service; fault-free from the perspective of the underlying built-in software functional failures previously described with regards to the prior art crawlers.

In another embodiment, the crawler is integrated into a Grandcentral station framework as a prototype.

**BRIEF DESCRIPTION OF THE FIGURES**

The subject matter which is regarded as the invention is particularly pointed out and distinctly claimed in the claims at the conclusion of the specification. The foregoing and other objects, features, and advantages of the invention will be apparent from the following detailed description taken in conjunction with the accompanying drawings.

FIG. 1 is a system level overview of a typical information processing network within which the present invention may be practiced.

FIG. 2 is a flow diagram that illustrates the web crawling and web browsing process found in the prior art.

FIG. 3a is a block diagram that depicts the system architecture for an enhanced browser based crawler.

FIG. 3b is a flow diagram that illustrates the enhanced browser based web crawling functional overview.

FIG. 4 is a flow diagram that illustrates the processing steps executed in a Pool Manager subroutine to control a visited pool and a URL pool.

FIG. 5 is a flow diagram that illustrates the processing steps executed in a Page Gatherer subroutine to gather contents of HTML pages.

FIG. 6 is a flow diagram that illustrates the processing steps executed in a Page Renderer subroutine to create an in-memory representation of a HTML page layout.

**EXPRESS MAIL NO.: EL563154951US**

FIG. 7 is a flow diagram that illustrates the processing steps executed in a Page Extractor subroutine to extract embedded information from documents.

FIG. 8 is a flow diagram that illustrates the processing steps executed in a Page Summarizer subroutine to summarize data into metadata for forwarding to an application.

5

**DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS**

It is important to note that these embodiments are only examples of the many advantageous uses of the innovative teachings herein. In general, statements made in the specification of the present application do not necessarily limit any of the various claimed inventions. Moreover, some statements may apply to some inventive features but not to others. In general, unless otherwise indicated, singular elements may be in the plural and vice versa with no loss of generality.

In the drawing like numerals refer to like parts through several views.

Glossary of Terms Used in this Disclosure

- Crawler - A program that automatically explores the World Wide Web by retrieving a document and recursively retrieving some or all the documents that are linked to it. The crawler visits web sites and reads their pages and other information in order to create entries for a search engine index. The major search engines on the Web all have such a program, which is also known as a "crawler" or a "bot." Crawlers or spiders are typically programmed to visit sites that have been submitted by their owners as new or updated. Entire sites or specific pages can be selectively visited and indexed. Spiders are called spiders because they usually visit many sites in parallel at the same time, their "legs" spanning a large area of the "web." Spiders can crawl through a site's pages in several ways. One way is to follow all the hypertext links in each page until all the pages have been read. The spider for the AltaVista search engine and its Web site is called Scooter. Scooter adheres to the rules of politeness for Web spiders that are specified in the Standard for Robot

## EXPRESS MAIL NO.: EL563154951US

Exclusion (SRE). It asks each server which files should be excluded from being indexed. It does not (or can not) go through firewalls. And it uses a special algorithm for waiting between successive server requests so that it doesn't affect response time for other users.

- 5 • Dictionary - A database of context-related terms.
- HTML (Hypertext Markup Language) - A standard language for attaching presentation and linking attributes to informational content within documents. During a document authoring stage, HTML "tags" are embedded within the informational content of the document. When the web document (or "HTML document") is subsequently transmitted by a web server to a web browser, the tags are interpreted by the browser and used to parse and display the document. In addition to specifying how the web browser is to display the document, HTML tags can be used to create hyperlinks to other web documents.
- 10 • Internet - A collection of interconnected public and private computer networks that are linked together with routers by a set of standards protocols to form a global, distributed network.
- Search engine - A remotely accessible World Wide Web tool that allows users to conduct keyword searches for information on the Internet. Typically, a search engine uses a spider (also called a "crawler" or a "bot") that goes to every page or representative pages on every Web site that wants to be searchable and reads it, using hypertext links on each page to discover and read a site's other pages. The spider creates a huge index (sometimes called a "catalog") from the pages that have been read. See crawler definition above.
- Server - A software program or a computer that responds to requests from a web browser by returning ("serving") web documents.
- 25 • URL (Uniform Resource Locator) - A unique address that fully specifies the location of a content object on the Internet. The general format of a URL is protocol://server-address/path/filename.
- Web browser - A software program that allows users to request and read hypertext

**EXPRESS MAIL NO.: EL563154951US**

documents. The browser gives some means of viewing the contents of web documents and of navigating from one document to another. Popular examples are Microsoft's Internet Explorer or Netscape's Navigator.

- Web document or page - A collection of data available on the World Wide Web and identified by a URL. In the simplest, most common case, a web page is a file written in HTML and stored on a web server. It is possible for the server to generate pages dynamically in response to a request from the user. A web page can be in any format that the browser or a helper application can display. The format is transmitted as part of the headers of the response as a MIME type, e.g. "text/html", "image/gif". An HTML web
- page will typically refer to other web pages and Internet resources by including hypertext links.
- Web Site - A database or other collection of inter-linked hypertext documents ("web documents" or "web pages") and associated data entities, which is accessible via a computer network, and which forms part of a larger, distributed informational system such as the WWW. In general, a web site corresponds to a particular Internet domain name, and includes the content of a particular organization. Other types of web sites may include, for example, a hypertext database of a corporate "intranet" (i.e., an internal network which uses standard Internet protocols), or a site of a hypertext system that uses document retrieval protocols other than those of the WWW.
- World Wide Web (WWW): An Internet client - server hypertext distributed information retrieval system.

Exemplary Embodiment-System Architecture for Enhanced Browser Based Crawler

The enhanced crawler will be described with the use of FIGs. 3 to 8 to describe the

**EXPRESS MAIL NO.: EL563154951US**

improved crawler services. The enhanced crawler is depicted in FIG. 3a and it comprises the following architecture:

- 1) URL Pool - 302a
- 2) Visited Pool - 304a
- 3) Pool Manager - 306a
- 4) Page Gatherer - 308a
- 5) Page Renderer - 310a
- 6) Resource Cache - 312a
- 7) Page Extractor - 314a
- 8) Page Summarizer -316a

FIG.3b is a flow diagram that illustrates an overview of the processing steps executed in accordance with the principles of this invention demonstrating the enhanced web crawling services of this invention. Essentially, the Enhanced Browser Based Crawler acts similarly to a web browser after retrieving the initially requested document (302b). It then loads additional or included documents (304b) as needed or required (e.g. inline-frames, frames, images, applets, audio, video, or equivalents). The Crawler then executes client side script or code (306b) and produces the final HTML markup (308b). This final HTML markup is ordinarily used for the rendering for user presentation process. However, unlike a web browser this invention does not render the composed document for viewing purposes. Rather it analyzes or summarizes it (310b), thereby extracting valuable metadata and other important information contained within the document. Finally, image data including images and image map are analyzed using optical character recognition (OCR) techniques (312b).

The following is a detailed functional description of the individual components that comprise the system and method for enhanced browser-based web crawling.

URL POOL 302a

As in other crawlers, a URL list or "Pool" stores URLs that are yet to be crawled by

the system. The pool is initialized with "seed URLs" that point to Web pages where crawling will begin. As pages are gathered from the Web, they are processed by the other components of this system. As new URLs are discovered by the Page Summarizer, they are given to the Pool Manager which inspects them, checking whether they exist in the Visited Pool, and if not, inserts them into the URL pool for subsequent crawling. URLs are removed from the URL Pool when they are scheduled for crawling by the Pool Manager. The crawler system is finished when the URL pool is empty and all other components are idle.

#### Visited Pool 304a

URLs that have been crawled are inserted into a Visited list or "Pool". This pool accumulates over time and will eventually contain all URLs gathered by the crawler system.

#### Pool Manager 306a

FIG. 4 is a flow diagram that illustrates the processing steps executed in a Pool Manager subroutine to control a visited pool and a URL pool. The Pool Manager is responsible for maintaining the contents of the URL Pool 302a and the Visited Pool 304a. The URL Pool is initialized with seed URLs through the Pool Manager 402. As pages are gathered from the Web, they are processed by the other components of this system. As new URLs 404 are discovered by the Page Summarizer, they are given to the Pool Manager which inspects them, checking whether they exist in the Visited Pool 406, and if not, inserts them into the URL pool 408 for subsequent crawling. The Pool Manager schedules a URL for crawling 410 and then the URL is crawled at the scheduled time 412. The URL is removed from the URL Pool 414 by the Pool Manager subsequent to the crawling and entered into the Visited Pool 416. The crawler system is finished when the URL pool is empty and all other components are idle.

#### Page Gatherer 308a

FIG. 5 is a flow diagram that illustrates the processing steps executed in a Page

**EXPRESS MAIL NO.: EL563154951US**

Gatherer subroutine to gather contents of HTML pages. The Pool Manager gives a URL to the Page Gatherer for gathering 502. The Page Gatherer issues an HTTP command to the Web server named in the URL 504 and subsequently receives the content of the HTML page 506. The contents of the page are passed on to the Page Renderer for rendering 508. A check is performed to see if there are more URLs that need to be gathered at 510; the process continues as long as there are URLs to be gathered.

Page Renderer 310a

FIG. 6 is a flow diagram that illustrates the processing steps executed in a Page Renderer subroutine to create an in-memory representation of a HTML page layout. The Page Renderer is a page processing engine that exists in a Web browser and is normally used for displaying a HTML page to the user. In the Enhanced Browser Based Crawler system, however, the Page Renderer is used as an intermediate component for extracting the contents of a HTML page. The enhanced crawler relies on the features of industrial-strength Page Renderers such as the one included in the publicly available Mozilla browser, also known as Netscape Navigator. These renderers implement several features that a crawler needs for effective data extraction, including:

- Handling of HTML text (tables, paragraphs, lists, or equivalents);
- Handling of inline GIF and JPEG graphics (buttons, banners, maps, or equivalents);
- Execution of JavaScript code;
- Processing of HTML frames;
- Potentially, execution of Java applet code.

The Page Renderer receives the contents of a HTML page from the Page Gatherer 602. It processes the contents by building an in-memory representation 604 of the layout of the page on a would-be user interface. These memory structures would normally communicate to a user interface component information on the manner of laying out text and graphics on the page. However, in this Enhanced Browser Based Crawler system, the memory

structures are passed on to the Page Extractor for processing 616 when no additional pages are needed for creation of the in-memory representation.

In the event more pages are needed 606 in order to build a full representation of the layout of a HTML page, the Page Renderer may have to request additional pages to be fetched from the Web 608. These additional pages can include child frames and in-line GIF and JPEG image files (and potentially Java applet code). These URL requests are given to the Page Gatherer which retrieves 610 them immediately. The visited URLs are inserted into the Visited Pool and the Resource Cache 612. Finally, a final representation is constructed 614 and passed to the page extractor 616.

#### Resource Cache 312a

The Resource Cache stores a copy of each child frame page and in-line GIF and JPEG image files. The Page Renderer may need them in computing the layout of several HTML pages, so it is efficient to keep them in a local store rather than fetching these several times. Note that every page is not cached, just those pages that are used as child frames and those image files that are used in-line in pages.

#### Page Extractor 314a

FIG. 7 is a flow diagram that illustrates the processing steps executed in a Page Extractor subroutine to extract embedded information from documents. This component is activated when the Page Renderer has finished rendering a HTML page. The Page Extractor has access to the memory structures of the Page Renderer 702. It first copies the contents of the text portion of the page into its own data structure 704, the Text Map. It then inspects in-line GIF and JPEG image references 706 and extracts the "alternate text" attributes 708 which typically describe the contents of an image (e.g. button or banner) in text form. These text attributes are also added 710 to the Text Map. Next, an optical character recognition engine is invoked 712 which analyzes all in-line GIF and JPEG images 714 and attempts to extract textual content from them. This text content is added to the Text Map 716.



Page Summarizer 316a

FIG. 8 is a flow diagram that illustrates the processing steps executed in a Page Summarizer subroutine to summarize data into metadata for forwarding to an application. The Page Summarizer receives the Text Map from the Page Extractor 802 and processes the content in an application-specific manner 804. For instance, a crawler for price data would apply data extraction patterns 806 to the Text Map and translate it into structured price data 808. URLs residing in the Text Map are given to 810 the Pool Manager for subsequent crawling. Extracted data and / or metadata are given to the application logic 812.

In this manner, an improved System and Method for Enhanced Browser Based Web Crawling has been described that overcomes the imperfections of the prior art. Now, web crawling will not be a faulty process that does not permit accurate retrieval of dynamic content and embedded image content. Rather, the use of browser technology integrated within the crawler in combination with the use of optical character recognition techniques disclosed herein allows for the accessing, retrieval and summarization of a whole and complete document free from any underlying software errors.

Discussion of Hardware and Software Implementation Options

The present invention, as would be known to one of ordinary skill in the art could be produced in hardware or software, or in a combination of hardware and software. The system, or method, according to the inventive principles as disclosed in connection with the preferred embodiment, may be produced in a single computer system having separate elements or means for performing the individual functions or steps described or claimed or one or more elements or means combining the performance of any of the functions or steps disclosed or claimed, or may be arranged in a distributed computer system, interconnected by any suitable means as would be known by one of ordinary skill in art.

According to the inventive principles as disclosed in connection with the preferred embodiment, the invention and the inventive principles are not limited to any particular kind of computer system but may be used with any general purpose computer, as would be

**EXPRESS MAIL NO.: EL563154951US**

known to one of ordinary skill in the art, arranged to perform the functions described and the method steps described. The operations of such a computer, as described above, may be according to a computer program contained on a medium for use in the operation or control of the computer, as would be known to one of ordinary skill in the art. The computer medium which may be used to hold or contain the computer program product, may be a fixture of the computer such as an embedded memory or may be on a transportable medium such as a disk, as would be known to one of ordinary skill in the art.

The invention is not limited to any particular computer program or logic or language, or instruction but may be practiced with any such suitable program, logic or language, or instructions as would be known to one of ordinary skill in the art. Without limiting the principles of the disclosed invention any such computing system can include, inter alia, at least a computer readable medium allowing a computer to read data, instructions, messages or message packets, and other computer readable information from the computer readable medium. The computer readable medium may include non-volatile memory, such as ROM, Flash memory, floppy disk, Disk drive memory, CD-ROM, and other permanent storage. Additionally, a computer readable medium may include, for example, volatile storage such as RAM, buffers, cache memory, and network circuits.

Furthermore, the computer readable medium may include computer readable information in a transitory state medium such as a network link and/or a network interface, including a wired network or a wireless network, that allow a computer to read such computer readable information.

What is claimed is:

CLAIMS

1. A method for browser-enhanced web crawling associated with a network of hub processing units coupled to a plurality of information processing units over a network, the method executed by a web crawler on a hub processing unit associated with the network comprising the steps of:

retrieving a document at an address;  
loading secondary documents;  
sending to one or more information processing units a browser side script to gather metadata; and performing the sub-steps of:  
producing a final HTML markup;  
analyzing and summarizing the final HTML markup to produce metadata.

2. The method as defined in claim 1, wherein the retrieving a document at an address step further comprises retrieving a document at an address selected from the group of addresses consisting of a nodal address, a network address, a URL and equivalents.

3. The method as defined in claim 1, wherein the analyzing and summarizing step further comprises analyzing and summarizing the whole and complete document.

4. The method as defined in claim 1, further comprising the step of analyzing any image data present in the document and any image data present in the documents utilizing optical character recognition techniques.

5. The method as defined in claim 1, wherein the step of loading secondary documents further comprises the loading of secondary documents including documents selected from the group of documents consisting of in-line frames, frames, images, image maps, applets, audio, video or equivalents.

**EXPRESS MAIL NO.: EL563154951US**

6. The method as defined in claim 4, wherein the step of analyzing any image data present in the document and any image date present in the documents utilizing optical character recognition techniques further comprises analyzing any images and image maps in the image data to produce text data.

7. The method as defined in claim 1, wherein the retrieving step further comprises performing the sub-steps of:

initializing a first list with seed values;

checking if there are any URLs to be processed and if there are, performing the secondary sub-steps of:

determining if a URL is in a second list; and if it is not in the second list; then performing the tertiary sub-steps of:

inserting the URL into the first list;

scheduling the URL for crawling;

crawling the URL when scheduled to do so;

removing the URL from the first list after the scheduled crawling;

entering the URL into the second list; and

repeating the checking step until there are no more URLs to be processed;

where if the determining step determines that the URL is in the second list then repeating the checking step until there are no more URLs to be processed.

8. The method as defined in claim 7, wherein the sub-step of initializing a first list with seed values further includes the list being a URL pool.

9. The method as defined in claim 7, wherein the sub-step of determining if a URL is in a second list further includes the second list being a visited pool.

**EXPRESS MAIL NO.: EL563154951US**

1 10. The method as defined in claim 7, wherein the tertiary sub-step of crawling  
2 further comprises the sub-steps of:  
3 issuing an HTTP command to a web server named in the URL;  
4 receiving contents of an HTML page as a result of the issued HTTP command;  
5 and  
6 passing on the contents of the HTML page to a Page Rendering subroutine.

1 11. The method as defined in claim 10, further including the sub-steps performed by  
2 the Page Rendering subroutine comprising:  
3 receiving the contents of the HTML page in the Page Rendering subroutine;  
4 building an in-memory representation of a Layout for the HTML page and if more  
5 data is needed to properly form the representation, then performing the sub-steps of:  
6 requesting additional web-based information;  
7 gathering this additional web-based information;  
8 inserting any URLs associated with this additional web-based information  
9 into the second list and a URL cache;  
10 building a final amended representation; and  
11 forwarding the final amended representation to an Extraction subroutine;  
12 wherein, if no more data is needed to properly form the in-memory representation, then  
13 forwarding the in-memory representation to the Extraction subroutine.

**EXPRESS MAIL NO.: EL563154951US**

12. The method as defined in claim 11, further including the sub-steps performed by the Page Extraction subroutine comprising:

- accessing a set of memory structures of the Page Renderer;
- copying a text portion of the structures into a text map;
- inspecting any in-line GIF and JPEG image references in the memory structures;
- extracting alternate text attributes;
- adding the alternate text attributes to a text map;
- invoking an optical character recognition engine;
- analyzing any in-line GIF and JPEG images using the optical character recognition engine for text content;
- extracting text content from the GIF and JPEG images;
- adding text content from the images to the text map; and
- forwarding the text map to a Page Summarizer subroutine.

13. The method as defined in claim 12, further including the sub-steps performed by the Page Summarizer subroutine comprising:

- receiving a text map from the Page Extractor subroutine;
- processing the text map in an application-specific manner;
- applying data extraction patterns to the text map;
- translating resultant data from the applying step;
- forwarding any URLs present in the text map to a manager subroutine; and
- forwarding any extracted data and metadata to application logic.

**EXPRESS MAIL NO.: EL563154951US**

1 14. A computer readable medium including programming instructions, the  
2 programming instructions including instructions for browser-enhanced web crawling  
3 associated with a network of hub processing units coupled to a plurality of information  
4 processing units over a network, the browser enhanced web crawling instructions on  
5 the computer readable medium comprising:

6 retrieving instructions for retrieving a document at an address;  
7 loading instructions for loading secondary documents;  
8 sending instructions for sending to one or more information processing units a  
9 browser side script to gather metadata;  
10 producing instructions for producing a final HTML markup;  
11 analyzing and summarizing instructions for analyzing and summarizing the final  
12 HTML markup to produce the final metadata.

1 15. The computer readable medium as defined in claim 14, wherein the retrieving  
2 instructions for retrieving a document at an address further comprises retrieving  
3 instructions for retrieving a document at an address selected from the group of  
4 addresses consisting of a nodal address, a network address, a URL and equivalents.

1 16. The computer readable medium as defined in claim 14, wherein the analyzing  
2 and summarizing instructions further comprise analyzing and summarizing instructions  
3 for analyzing and summarizing the whole and complete document.

1 17. The computer readable medium as defined in claim 14, further comprising image  
2 analyzing instructions for analyzing any image data present in the document and any  
3 image data present in the documents utilizing optical character recognition techniques.

**EXPRESS MAIL NO.: EL563154951US**

1 18. The computer readable medium as defined in claim 14, wherein the loading  
2 instructions for loading secondary documents further comprises loading instructions for  
3 loading of secondary documents including documents selected from the group of  
4 documents consisting of in-line frames, frames, images, image maps, applets, audio,  
5 video or equivalents.

1 19. The computer readable medium as defined in claim 17, wherein the analyzing  
2 instructions for analyzing any image data present in the document and any image data  
3 present in the documents utilizing optical character recognition techniques further  
4 comprises analyzing instructions for analyzing any images and image maps in the  
5 image data to produce text data.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65  
66  
67  
68  
69  
70  
71  
72  
73  
74  
75  
76  
77  
78  
79  
80  
81  
82  
83  
84  
85  
86  
87  
88  
89  
90  
91  
92  
93  
94  
95  
96  
97  
98  
99  
100  
101  
102  
103  
104  
105  
106  
107  
108  
109  
110  
111  
112  
113  
114  
115  
116  
117  
118  
119  
120  
121  
122  
123  
124  
125  
126  
127  
128  
129  
130  
131  
132  
133  
134  
135  
136  
137  
138  
139  
140  
141  
142  
143  
144  
145  
146  
147  
148  
149  
150  
151  
152  
153  
154  
155  
156  
157  
158  
159  
160  
161  
162  
163  
164  
165  
166  
167  
168  
169  
170  
171  
172  
173  
174  
175  
176  
177  
178  
179  
180  
181  
182  
183  
184  
185  
186  
187  
188  
189  
190  
191  
192  
193  
194  
195  
196  
197  
198  
199  
200  
201  
202  
203  
204  
205  
206  
207  
208  
209  
210  
211  
212  
213  
214  
215  
216  
217  
218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229  
230  
231  
232  
233  
234  
235  
236  
237  
238  
239  
240  
241  
242  
243  
244  
245  
246  
247  
248  
249  
250  
251  
252  
253  
254  
255  
256  
257  
258  
259  
260  
261  
262  
263  
264  
265  
266  
267  
268  
269  
270  
271  
272  
273  
274  
275  
276  
277  
278  
279  
280  
281  
282  
283  
284  
285  
286  
287  
288  
289  
290  
291  
292  
293  
294  
295  
296  
297  
298  
299  
300  
301  
302  
303  
304  
305  
306  
307  
308  
309  
310  
311  
312  
313  
314  
315  
316  
317  
318  
319  
320  
321  
322  
323  
324  
325  
326  
327  
328  
329  
330  
331  
332  
333  
334  
335  
336  
337  
338  
339  
340  
341  
342  
343  
344  
345  
346  
347  
348  
349  
350  
351  
352  
353  
354  
355  
356  
357  
358  
359  
360  
361  
362  
363  
364  
365  
366  
367  
368  
369  
370  
371  
372  
373  
374  
375  
376  
377  
378  
379  
380  
381  
382  
383  
384  
385  
386  
387  
388  
389  
390  
391  
392  
393  
394  
395  
396  
397  
398  
399  
400  
401  
402  
403  
404  
405  
406  
407  
408  
409  
410  
411  
412  
413  
414  
415  
416  
417  
418  
419  
420  
421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431  
432  
433  
434  
435  
436  
437  
438  
439  
440  
441  
442  
443  
444  
445  
446  
447  
448  
449  
450  
451  
452  
453  
454  
455  
456  
457  
458  
459  
460  
461  
462  
463  
464  
465  
466  
467  
468  
469  
470  
471  
472  
473  
474  
475  
476  
477  
478  
479  
480  
481  
482  
483  
484  
485  
486  
487  
488  
489  
490  
491  
492  
493  
494  
495  
496  
497  
498  
499  
500  
501  
502  
503  
504  
505  
506  
507  
508  
509  
510  
511  
512  
513  
514  
515  
516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539  
540  
541  
542  
543  
544  
545  
546  
547  
548  
549  
550  
551  
552  
553  
554  
555  
556  
557  
558  
559  
560  
561  
562  
563  
564  
565  
566  
567  
568  
569  
570  
571  
572  
573  
574  
575  
576  
577  
578  
579  
580  
581  
582  
583  
584  
585  
586  
587  
588  
589  
590  
591  
592  
593  
594  
595  
596  
597  
598  
599  
600  
601  
602  
603  
604  
605  
606  
607  
608  
609  
610  
611  
612  
613  
614  
615  
616  
617  
618  
619  
620  
621  
622  
623  
624  
625  
626  
627  
628  
629  
630  
631  
632  
633  
634  
635  
636  
637  
638  
639  
640  
641  
642  
643  
644  
645  
646  
647  
648  
649  
650  
651  
652  
653  
654  
655  
656  
657  
658  
659  
660  
661  
662  
663  
664  
665  
666  
667  
668  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701  
702  
703  
704  
705  
706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
720  
721  
722  
723  
724  
725  
726  
727  
728  
729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755  
756  
757  
758  
759  
760  
761  
762  
763  
764  
765  
766  
767  
768  
769  
770  
771  
772  
773  
774  
775  
776  
777  
778  
779  
780  
781  
782  
783  
784  
785  
786  
787  
788  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809  
810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863  
864  
865  
866  
867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917  
918  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971  
972  
973  
974  
975  
976  
977  
978  
979  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
1000  
1001  
1002  
1003  
1004  
1005  
1006  
1007  
1008  
1009  
1010  
1011  
1012  
1013  
1014  
1015  
1016  
1017  
1018  
1019  
1020  
1021  
1022  
1023  
1024  
1025  
1026  
1027  
1028  
1029  
1030  
1031  
1032  
1033  
1034  
1035  
1036  
1037  
1038  
1039  
1040  
1041  
1042  
1043  
1044  
1045  
1046  
1047  
1048  
1049  
1050  
1051  
1052  
1053  
1054  
1055  
1056  
1057  
1058  
1059  
1060  
1061  
1062  
1063  
1064  
1065  
1066  
1067  
1068  
1069  
1070  
1071  
1072  
1073  
1074  
1075  
1076  
1077  
1078  
1079  
1080  
1081  
1082  
1083  
1084  
1085  
1086  
1087  
1088  
1089  
1090  
1091  
1092  
1093  
1094  
1095  
1096  
1097  
1098  
1099  
1100  
1101  
1102  
1103  
1104  
1105  
1106  
1107  
1108  
1109  
1110  
1111  
1112  
1113  
1114  
1115  
1116  
1117  
1118  
1119  
1120  
1121  
1122  
1123  
1124  
1125  
1126  
1127  
1128  
1129  
1130  
1131  
1132  
1133  
1134  
1135  
1136  
1137  
1138  
1139  
1140  
1141  
1142  
1143  
1144  
1145  
1146  
1147  
1148  
1149  
1150  
1151  
1152  
1153  
1154  
1155  
1156  
1157  
1158  
1159  
1160  
1161  
1162  
1163  
1164  
1165  
1166  
1167  
1168  
1169  
1170  
1171  
1172  
1173  
1174  
1175  
1176  
1177  
1178  
1179  
1180  
1181  
1182  
1183  
1184  
1185  
1186  
1187  
1188  
1189  
1190  
1191  
1192  
1193  
1194  
1195  
1196  
1197  
1198  
1199  
1200  
1201  
1202  
1203  
1204  
1205  
1206  
1207  
1208  
1209  
1210  
1211  
1212  
1213  
1214  
1215  
1216  
1217  
1218  
1219  
1220  
1221  
1222  
1223  
1224  
1225  
1226  
1227  
1228  
1229  
1230  
1231  
1232  
1233  
1234  
1235  
1236  
1237  
1238  
1239  
1240  
1241  
1242  
1243  
1244  
1245  
1246  
1247  
1248  
1249  
1250  
1251  
1252  
1253  
1254  
1255  
1256  
1257  
1258  
1259  
1260  
1261  
1262  
1263  
1264  
1265  
1266  
1267  
1268  
1269  
1270  
1271  
1272  
1273  
1274  
1275  
1276  
1277  
1278  
1279  
1280  
1281  
1282  
1283  
1284  
1285  
1286  
1287  
1288  
1289  
1290  
1291  
1292  
1293  
1294  
1295  
1296  
1297  
1298  
1299  
1300  
1301  
1302  
1303  
1304  
1305  
1306  
1307  
1308  
1309  
1310  
1311  
1312  
1313  
1314  
1315  
1316  
1317  
1318  
1319  
1320  
1321  
1322  
1323  
1324  
1325  
1326  
1327  
1328  
1329  
1330  
1331  
1332  
1333  
1334  
1335  
1336  
1337  
1338  
1339  
1340  
1341  
1342  
1343  
1344  
1345  
1346  
1347  
1348  
1349  
1350  
1351  
1352  
1353  
1354  
1355  
1356  
1357  
1358  
1359  
1360  
1361  
1362  
1363  
1364  
1365  
1366  
1367  
1368  
1369  
1370  
1371  
1372  
1373  
1374  
1375  
1376  
1377  
1378  
1379  
1380  
1381  
1382  
1383  
1384  
1385  
1386  
1387  
1388  
1389  
1390  
1391  
1392  
1393  
1394  
1395  
1396  
1397  
1398  
1399  
1400  
1401  
1402  
1403  
1404  
1405  
1406  
1407  
1408  
1409  
1410  
1411  
1412  
1413  
1414  
1415  
1416  
1417  
1418  
1419  
1420  
1421  
1422  
1423  
1424  
1425  
1426  
1427  
1428  
1429  
1430  
1431  
1432  
1433  
1434  
1435  
1436  
1437  
1438  
1439  
1440  
1441  
1442  
1443  
1444  
1445  
1446  
1447  
1448  
1449  
1450  
1451  
1452  
1453  
1454  
1455  
1456  
1457  
1458  
1459  
1460  
1461  
1462  
1463  
1464  
1465  
1466  
1467  
1468  
1469  
1470  
1471  
1472  
1473  
1474  
1475  
1476  
1477  
1478  
1479  
1480  
1481  
1482  
1483  
1484  
1485  
1486  
1487  
1488  
1489  
1490  
1491  
1492  
1493  
1494  
1495  
1496  
1497  
1498  
1499  
1500  
1501  
1502  
1503  
1504  
1505  
1506  
1507  
1508  
1509  
1510  
1511  
1512  
1513  
1514  
1515  
1516  
1517  
1518  
1519  
1520  
1521  
1522  
1523  
1524  
1525  
1526  
1527  
1528  
1529  
1530  
1531  
1532  
1533  
1534  
1535  
1536  
1537  
1538  
1539  
1540  
1541  
1542  
1543  
1544  
1545  
1546  
1547  
1548  
1549  
1550  
1551  
1552  
1553  
1554  
1555  
1556  
1557  
1558  
1559  
1560  
1561  
1562  
1563  
1564  
1565  
1566  
1567  
1568  
1569  
1570  
1571  
1572  
1573  
1574  
1575  
1576  
1577  
1578  
1579  
1580  
1581  
1582  
1583  
1584  
1585  
1586  
1587  
1588  
1589  
1590  
1591  
1592  
1593  
1594  
1595  
1596  
1597  
1598  
1599  
1600  
1601  
1602  
1603  
1604  
1605  
1606  
1607  
1608  
1609  
1610  
1611  
1612  
1613  
1614  
1615  
1616  
1617  
1618  
1619  
1620  
1621  
1622  
1623  
1624  
1625  
1626  
1627  
1628  
1629  
1630  
1631  
1632  
1633  
1634  
1635  
1636  
1637  
1638  
1639  
1640  
1641  
1642  
1643  
1644  
1645  
1646  
1647  
1648  
1649  
1650  
1651  
1652  
1653  
1654  
1655  
1656  
1657  
1658  
1659  
1660  
1661  
1662  
1663  
1664  
1665  
1666  
1667  
1668  
1669  
1670  
1671  
1672  
1673  
1674  
1675  
1676  
1677  
1678  
1679  
1680  
1681  
1682  
1683  
1684  
1685  
1686  
1687  
1688  
1689  
1690  
1691  
1692  
1693  
1694  
1695  
1696  
1697  
1698  
1699  
1700  
1701  
1702  
1703  
1704  
1705  
1706  
1707  
1708  
1709  
1710  
1711  
1712  
1713  
1714  
1715  
1716  
1717  
1718  
1719  
1720  
1721  
1722  
1723  
1724  
1725  
1726  
1727  
1728  
1729  
1730  
1731  
1732  
1733  
1734  
1735  
1736  
1737  
1738  
1739  
1740  
1741  
1742  
1743  
1744  
1745  
1746  
1747  
1748  
1749  
1750  
1751  
1752  
1753  
1754  
1755  
1756  
1757  
1758  
1759  
1760  
1761  
1762  
1763  
1764  
1765  
1766  
1767  
1768  
1769  
1770  
1771  
1772  
1773  
1774  
1775  
1776  
1777  
1778  
1779  
1780  
1781  
1782  
1783  
1784  
1785  
1786  
1787  
1788  
1789  
1790  
1791  
1792  
1793  
1794  
1795  
1796  
1797  
1798  
1799  
1800  
1801  
1802  
1803  
1804  
1805  
1806  
1807  
1808  
1809  
1810  
1811  
1812  
1813  
1814  
1815  
1816  
1817  
1818  
1819  
1820  
1821  
1822  
1823  
1824  
1825  
1826  
1827  
1828  
1829  
1830  
1831  
1832  
1833  
1834  
1835  
1836  
1837  
1838  
1839  
1840  
1841  
1842  
1843  
1844  
1845  
1846  
1847  
1848  
1849  
1850  
1851  
1852  
1853  
1854  
1855  
1856  
1857  
1858  
1859  
1860  
1861  
1862  
1863  
1864  
1865  
1866  
1867  
1868  
1869  
1870  
1871  
1872  
1873  
1874  
1875  
1876  
1877  
1878  
1879  
1880  
1881  
1882  
1883  
1884  
1885  
1886  
1887  
1888  
1889  
1890  
1891  
1892  
1893  
1894  
1895  
1896  
1897  
1898  
1899  
1900  
1901  
1902  
1903  
1904  
1905  
1906  
1907  
1908  
1909  
1910  
1911  
1912  
1913  
1914  
1915  
1916  
1917  
1918  
1919  
1920  
1921  
1922  
1923  
1924  
1925  
1926  
1927  
1928  
1929  
1930  
1931  
1932  
1933  
1934  
1935  
1936  
1937  
1938  
1939  
1940  
1941  
1942  
1943  
1944  
1945  
1946  
1947  
1948  
1949  
1950  
1951  
1952  
1953  
1954  
1955  
1956  
1957  
1958  
1959  
1960  
1961  
1962  
1963  
1964  
1965  
1966  
1967  
1968  
1969  
1970  
1971  
1972  
1973  
1974  
1975  
1976  
1977  
1978  
1979  
1980  
1981  
1982  
1983  
1984  
1985  
1986  
1987  
1988  
1989  
1990  
1991  
1992  
1993  
1994  
1995  
1996  
1997  
1998  
1999  
2000  
2001  
2002  
2003  
2004  
2005  
2006  
2007  
2008  
2009  
2010  
2011  
2012  
2013  
2014  
2015  
2016  
2017  
2018  
2019  
2020  
2021  
2022  
2023  
2024  
2025  
2026  
2027  
2028  
2029  
2030  
2031  
2032  
2033  
2034  
2035  
2036  
2037  
2038  
2039  
2040  
2041  
2042  
2043  
2044  
2045  
2046  
2047  
2048  
2049  
2050  
2051  
2052  
2053  
2054  
2055  
2056  
2057  
2058  
2059  
2060  
2061  
2062  
2063  
2064  
2065  
2066  
2067  
2068  
2069  
2070  
2071  
2072  
2073  
2074  
2075  
2076  
2077  
2078  
2079  
2080  
2081  
2082  
2083  
2084  
2085  
2086  
2087  
2088  
2089  
2090  
2091  
2092  
2093  
2094  
2095  
2096  
2097  
2098  
2099  
2100  
2101  
2102  
2103  
2104  
2105  
2106  
2107  
2108  
2109  
2110  
2111  
2112  
2113  
2114  
2115  
2116  
2117  
2118  
2119  
2120  
2121  
2122  
2123  
2124  
2125  
2126  
2127  
2128  
2129  
2130  
2131  
2132  
2133  
2134  
2135  
2136  
2137  
2138  
2139  
2140  
2141  
2142  
2143  
2144  
2145  
2146  
2147  
2148  
2149  
2150  
2151  
2152  
2153  
2154  
2155  
2156  
2157  
2158  
2159  
2160  
2161  
2162  
2163  
2164  
2165  
2166  
2167  
2168  
2169  
2170  
2171  
2172  
2173  
2174  
2175  
2176  
2177  
2178  
2179  
2180  
2181  
2182  
2183  
2184  
2185  
2186  
2187  
2188  
2189  
2190  
2191  
2192  
2193  
2194  
2195  
2196  
2197  
2198  
2199  
22



[illegible]

20. A browser-enhanced web crawling unit associated with a network of a plurality of hub processing units coupled to a plurality of information processing units over a network, the browser enhanced web crawling unit on a hub processing unit comprising:

- a retrieval unit for retrieving a document at an address;
- a loader for loading secondary documents as required;
- an output for sending to one or more information processing units a browser side script to gather metadata;
- a producer for producing a final HTML markup; and
- a summarizer for analyzing and summarizing the final HTML markup to produce the final metadata.

**EXPRESS MAIL NO.: EL563154951US**

**ABSTRACT**

This invention pioneers an enhanced crawling mechanism and technique called "Enhanced Browser Based Web Crawling". It permits the fault-tolerant gathering of dynamic data documents on the World Wide Web (WWW). The Enhanced Browser Based Web Crawler technology of this invention is implemented by incorporating the intricate functionality of a web browser into the crawler engine so that documents are properly analyzed. Essentially, the Enhanced Browser Based Crawler acts similarly to a web browser after retrieving the initially requested document. It then loads additional or included documents as needed or required (e.g. inline-frames, frames, images, applets, audio, video, or equivalents.). The Crawler then executes client side script or code and produces the final HTML markup. This final HTML markup is ordinarily used for the rendering for user presentation process. However, unlike a web browser this invention does not render the composed document for viewing purposes. Rather it analyzes or summarizes it, thereby extracting valuable metadata and other important information contained within the document. Also, this invention introduces the integration of optical character recognition (OCR) techniques into the crawler architecture. The reason for this is to enable the web crawler summarization process to properly summarize image content (e.g. GIF, JPEG or an equivalent) without errors.

110-A00-006v3.wpd

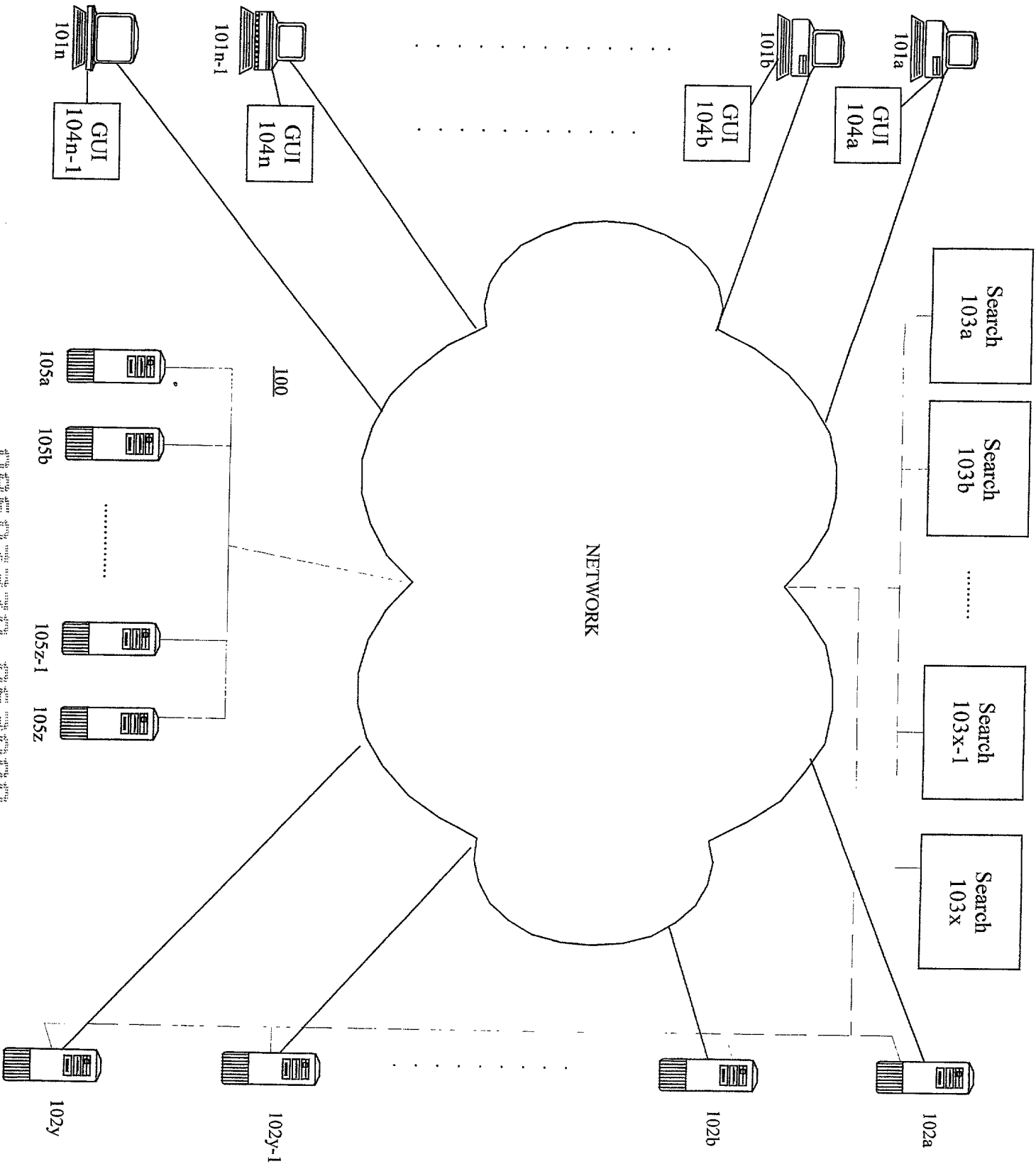


Fig. 1

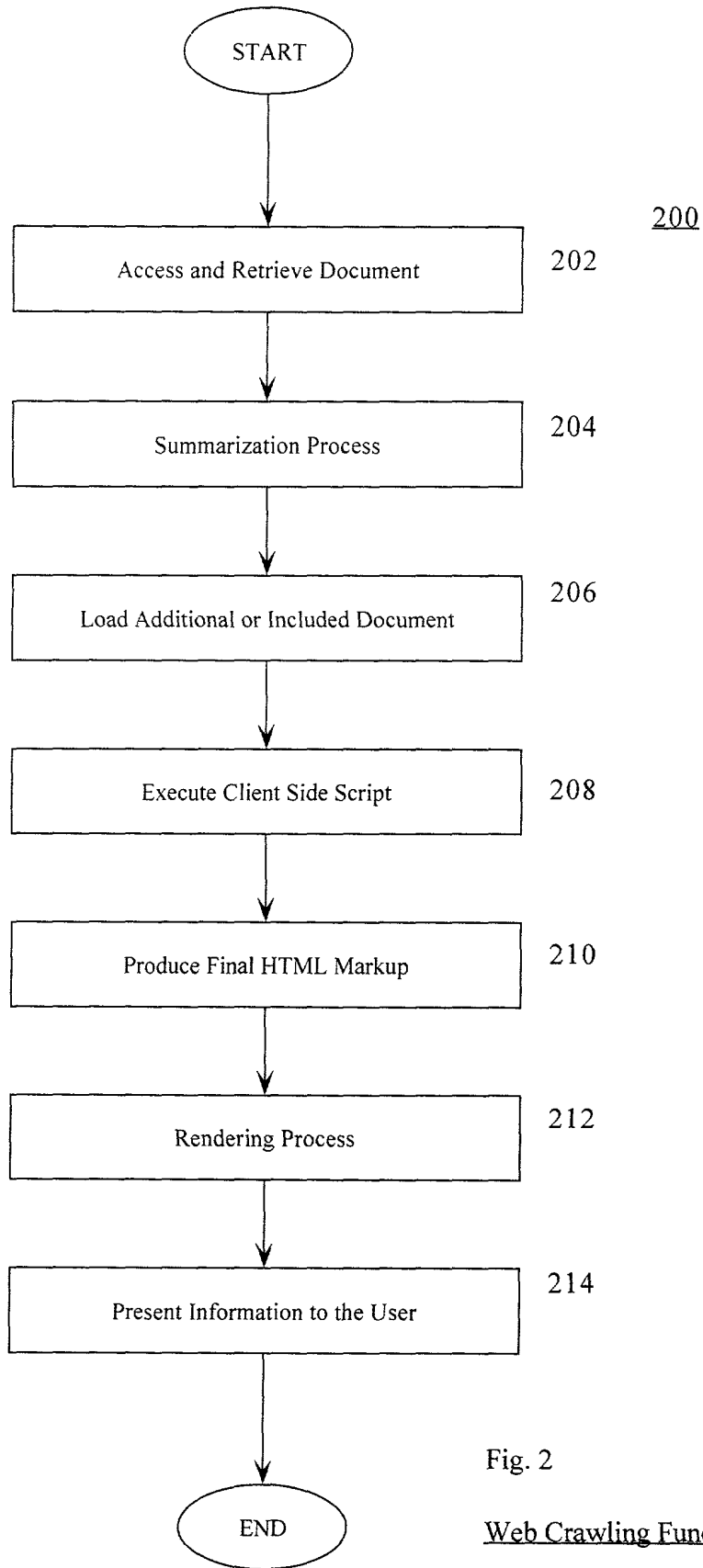


Fig. 2

Web Crawling Functional Overview

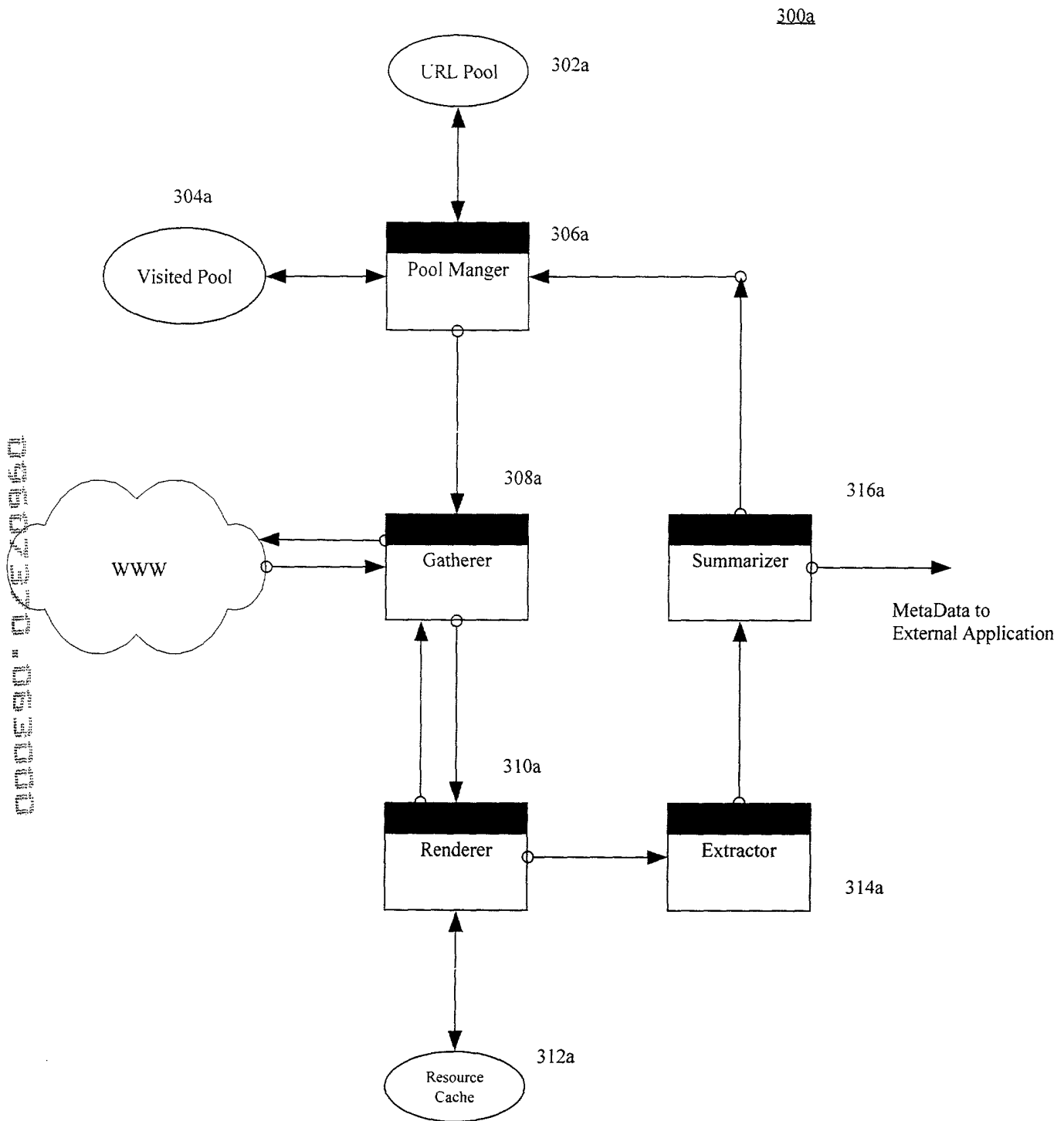


Fig. 3a System Architecture for Enhanced Browser-BasedCrawler

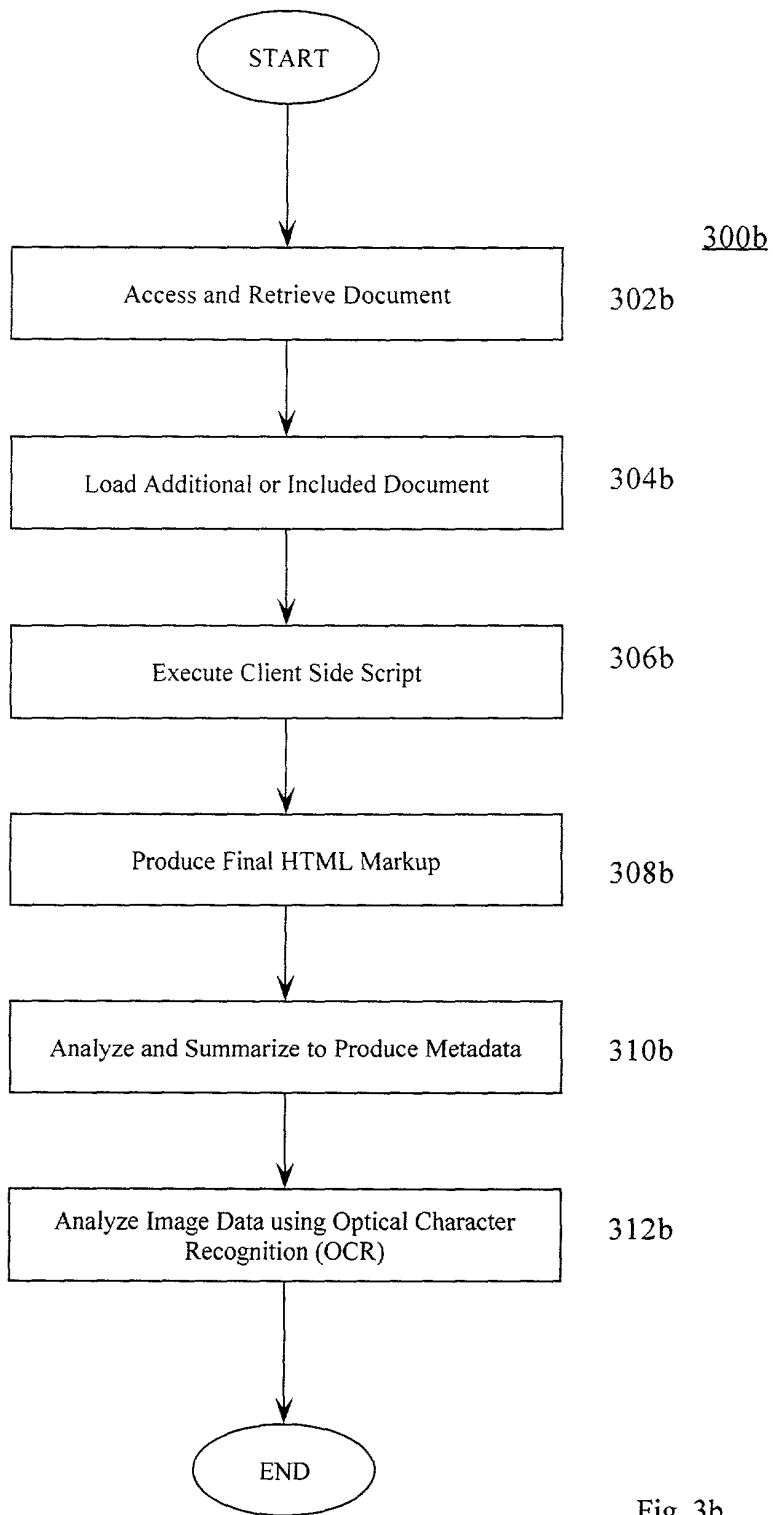


Fig. 3b

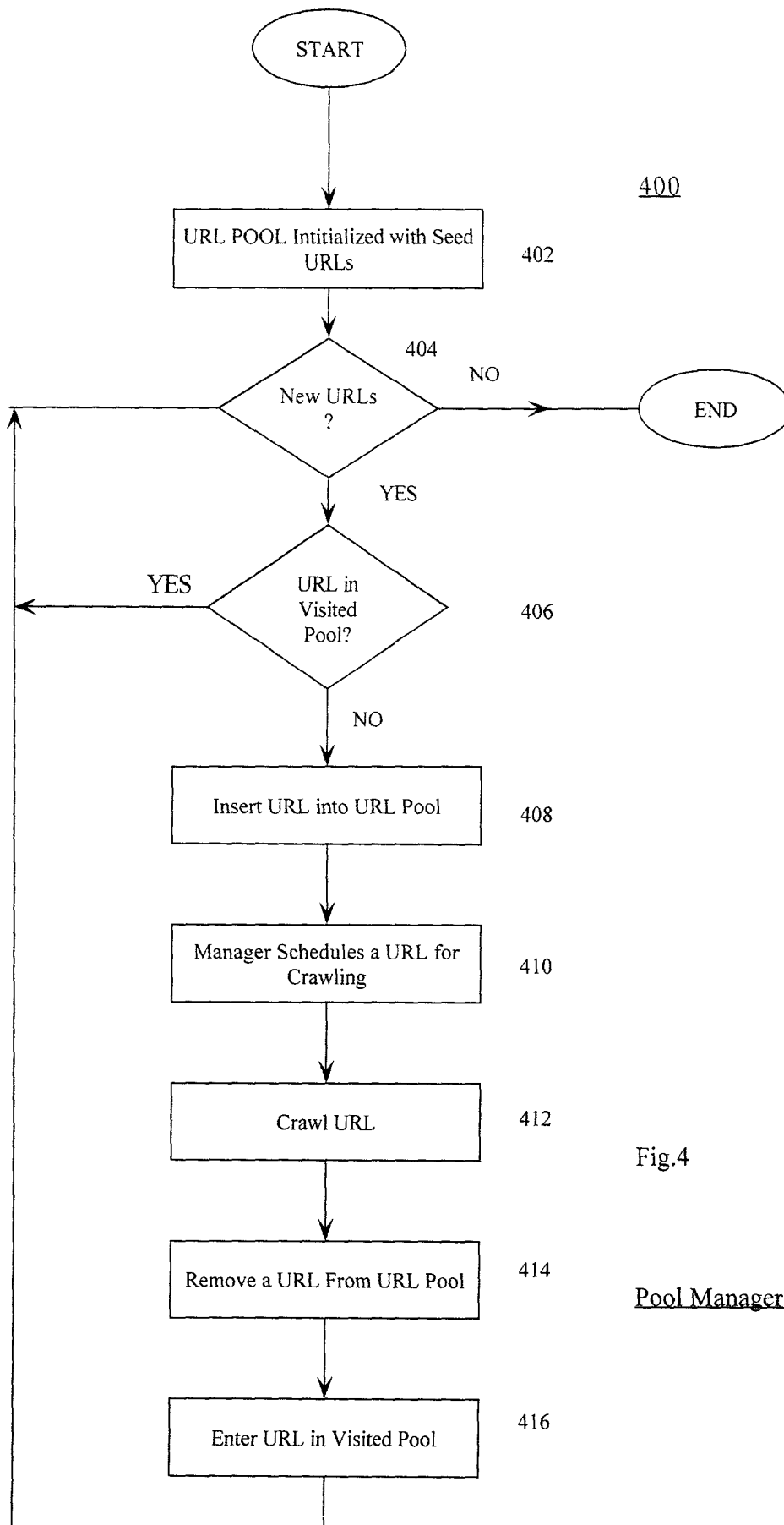
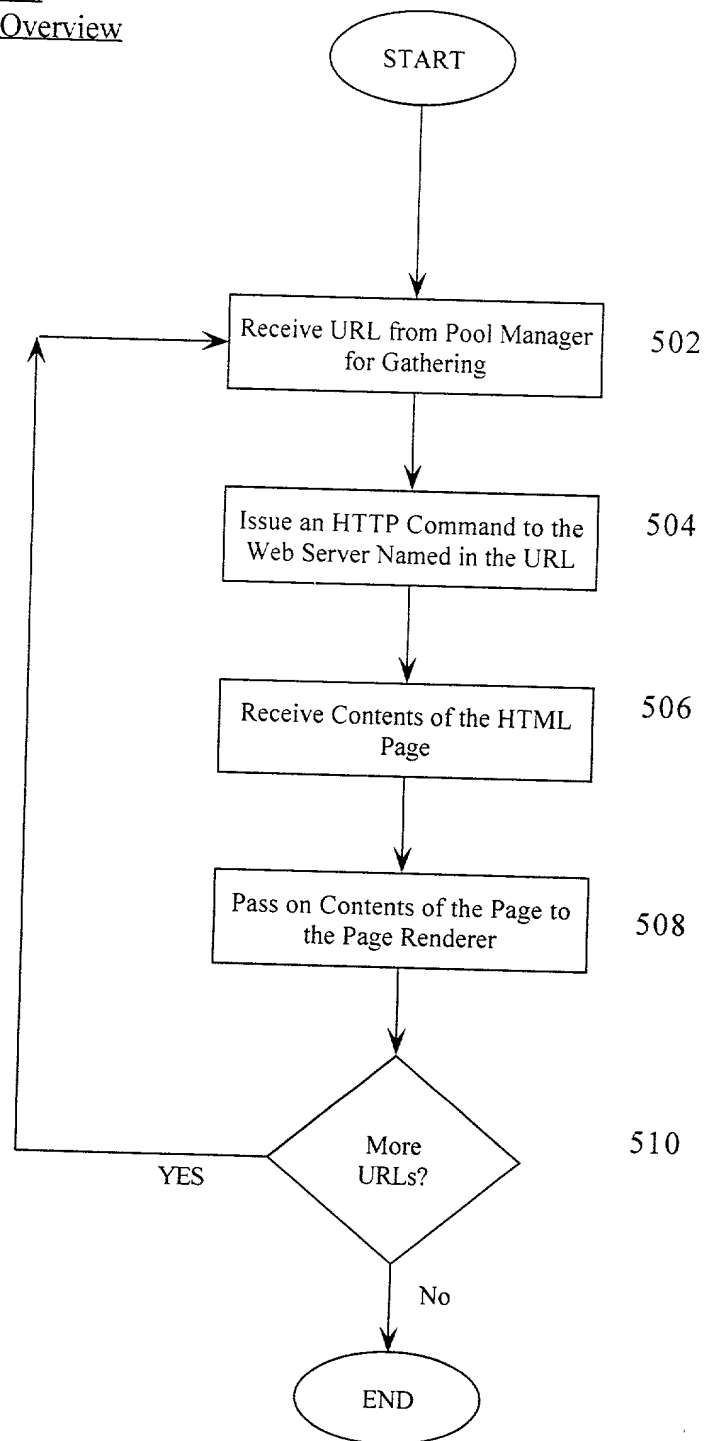


Fig.4

Pool Manager Operation

Page Gatherer  
Functional Overview



500

Fig. 5



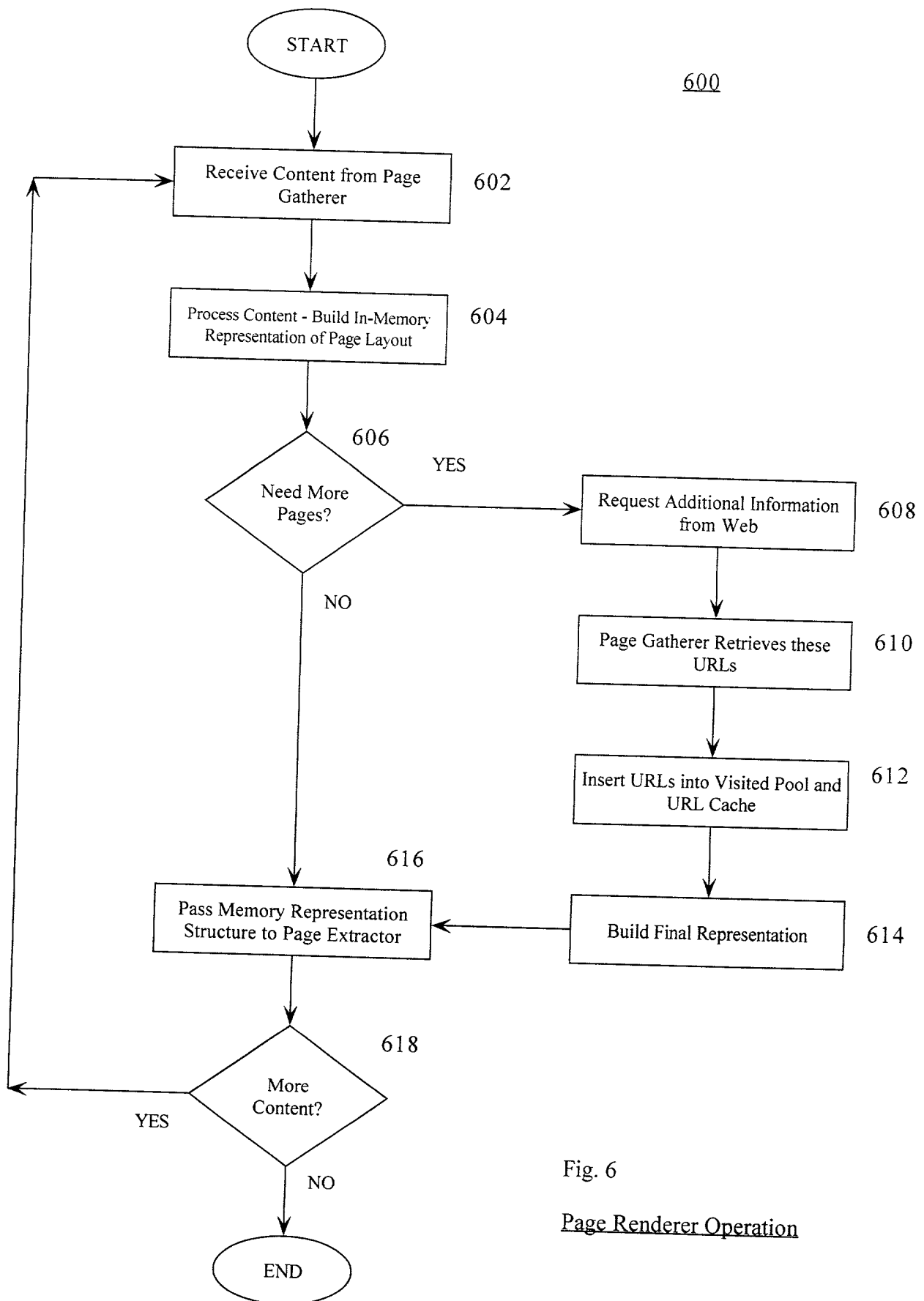
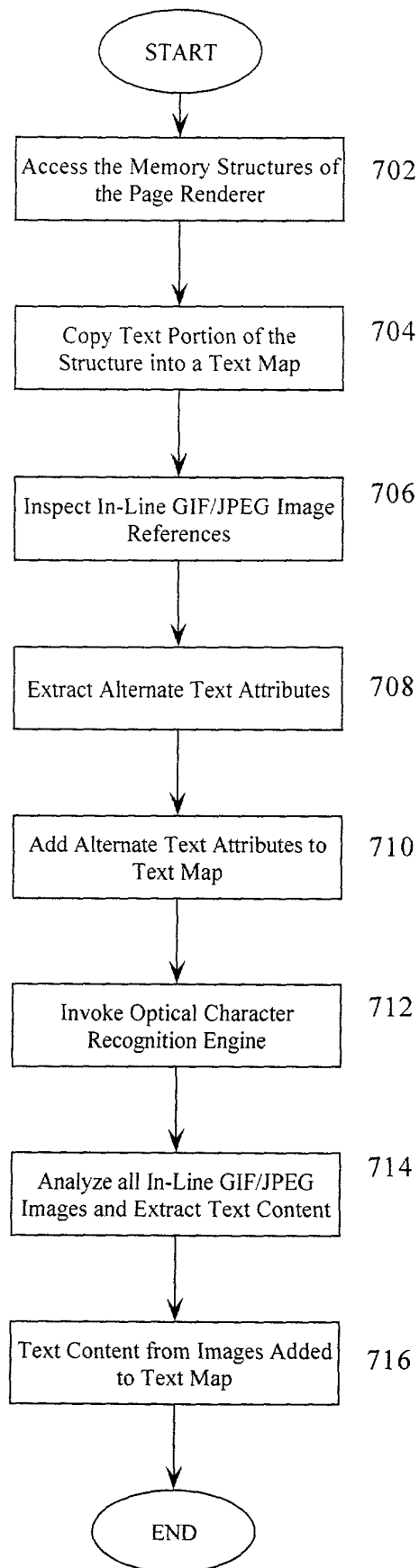


Fig. 6

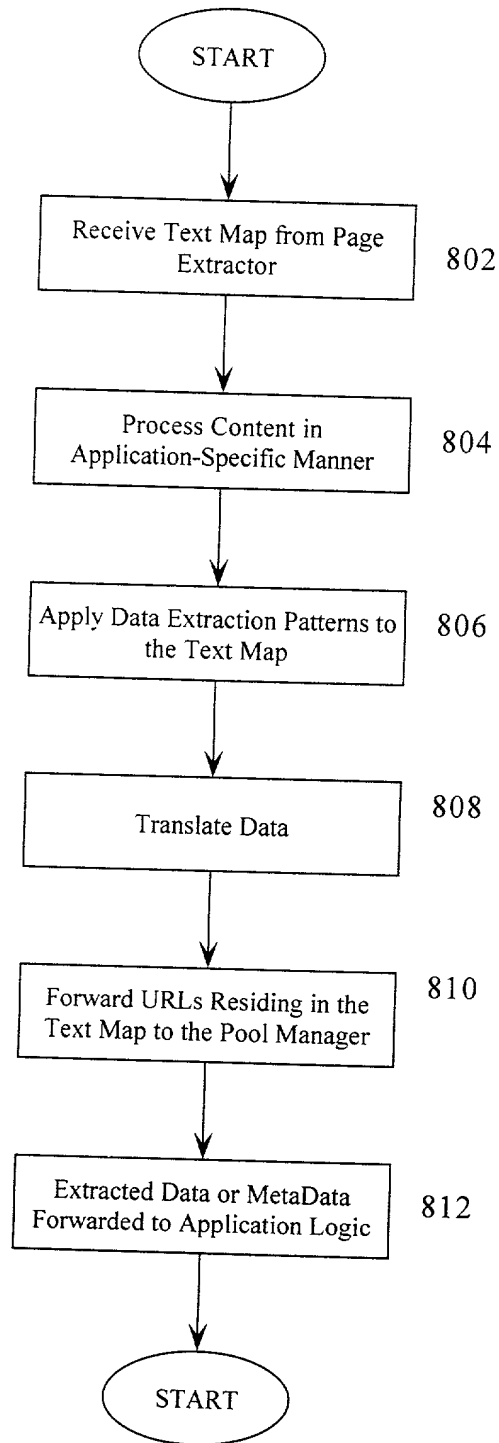
Page Renderer Operation

Page Extractor Operation



700

Fig.7



800

Fig. 8

Page Summarizer  
Operation

## DECLARATION AND POWER OF ATTORNEY FOR PATENT APPLICATION

As a below named inventor, I hereby declare that:

My residence, post office address and citizenship are as stated below next to my name;

I believe I am an original, first and joint inventor (if plural names are listed below) of the subject matter which is claimed and for which a patent is sought on the invention entitled

### SYSTEM AND METHOD FOR ENHANCED BROWSER-BASED WEB CRAWLING

the specification of which: (check one)

XXX is attached hereto.

\_\_\_\_\_ was filed on \_\_\_\_\_  
under Attorney's Docket Number \_\_\_\_\_  
as Application Serial No. \_\_\_\_\_  
and was amended on \_\_\_\_\_ (if applicable).

I hereby state that I have reviewed and understand the contents of the above identified specification, including the claims, as amended by any amendment referred to above.

I acknowledge the duty to disclose information which is material to the patentability of this application in accordance with 37 CFR 1.56.

I hereby claim the benefit of foreign priority under 35 USC 119 of any foreign application(s) for patent or inventor's certificate listed below and have also identified below any foreign application for patent or inventor's certificate having a filing date before that of the application the priority of which is claimed:

Prior Foreign Application(s):

Priority Claimed

\_\_\_\_\_  
(Number)

\_\_\_\_\_  
(Country)

\_\_\_\_\_  
(Filing Date)

\_\_\_ Yes \_\_\_ No

I hereby claim the benefit of United States priority under 35 USC 120 of any United States application(s) listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in a listed prior United States application in the manner provided by the first paragraph of 35 USC 112, I acknowledge the duty to disclose information material to the patentability of this application as defined in 37 CFR 1.56 which occurred between the filing date of the prior application and the national or PCT international filing date of this application:

\_\_\_\_\_  
(Application Serial #)

\_\_\_\_\_  
(Filing Date)

\_\_\_\_\_  
(Status)

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under 18 USC 1001 and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

POWER OF ATTORNEY: As a named inventor, I hereby appoint the following attorneys and/or agents to prosecute this application and transact all business in the Patent and Trademark Office connected therewith.

Khanh Q. Tran	Reg. No. 41,352	Jon A. Gibbons	Reg. No. 37,333
Thomas R. Berthold	Reg. No. 28,689	Stephen C. Bongini	Reg. No. 40,917
Richard M. Ludwin	Reg. No. 33,010	Jose Gutman	Reg. No. 35,171
Marc D. McSwain	Reg. No. 44,929	Martin Fleit	Reg. No. 16,900
Alison D. Mortinger	Reg. No. 39,306	Robert C. Kain	Reg. No. 30,648
		Philip Premysler	Reg. No. 43,015

Send correspondence to Jon A. Gibbons, Fleit, Kain, Gibbons, Gutman & Bongini P.L., 4400 N. Federal Highway, Suite 32, Boca Raton, Florida 33431 and direct all telephone calls to Jon A. Gibbons (561) 417-9477.

---

FULL NAME OF INVENTOR: Reiner KRAFT

INVENTOR'S SIGNATURE:  DATE: 6/27/2000

RESIDENCE: 9406 Wetsand Court, Gilroy, California 95020

CITIZENSHIP: Germany

POST OFFICE ADDRESS: Same as above

---

FULL NAME OF INVENTOR: Jussi P. MYLLYMAKI

INVENTOR'S SIGNATURE:  DATE: 6/27/2000

RESIDENCE: 395 Palm Ridge Lane, San Jose, California 95123

CITIZENSHIP: Finland

POST OFFICE ADDRESS: Same as above